AN ACCURATE AND EFFICIENT METHOD FOR RECONSTRUCTION OF 3D FACES FROM STEREO IMAGES

BY

VUONG VAN LE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Professor Thomas S. Huang

ABSTRACT

In this thesis, we introduce a novel algorithm for reconstructing the 3D shape and texture model of human faces from two stereo images which are captured from calibrated cameras. Our approach works in a sparse to dense manner: we first build a coarse shape estimation based on 3D keypoints, and then use a linear morphable model to efficiently match the detailed shape and texture.

The features used for the fitting processes are selected with the guidance of the quantitative evaluation of a state-of-the-art reconstruction algorithm. In our new direct evaluation method, the reconstructed 3D faces are first aligned to the ground truth and then the shape difference between the two 3D faces is described by signal-to-noise ratio and error maps illustrating the reconstruction errors on corresponding vertices. This local error information will be used to resample the reference frame whose vertices' coordinates stack up to be the feature vectors for face fitting.

Compared with the previous works, our algorithm can reconstruct the 3D face shape at a speed comparable with that of the fastest algorithm available, but gives a higher accuracy. It can also recover the more complete and realistic looking texture. Our results show that the new algorithm possesses significant characteristics of a 3D face model reconstruction system, and is especially useful for face recognition and animation applications in practice.

To my family and friends

ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Thomas S. Huang, for his great guidance and support. I would like to thank my labmates: Yuxiao Hu, Jason Xun Xu, Hao Tang, and Liangliang Cao for their advice, collaboration and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

3D face reconstruction from 2D images is an important problem in computer vision. In the past few decades, many approaches have been proposed, including 3D from stereo [1], 3D Morphable Model based methods [2], structure from motion [3] and shape from shading techniques [4].

Among these, 3D Morphable Model (3DMM) based algorithms have attracted more and more attention in recent years. Vetter et al. proposed a family of 3D fitting algorithms [2, 5, 6] to recover the facial shape and texture parameters of the 3DMM from the appearance features. Although being able to give accurate shape and controllable texture models, these methods tend to extract complicated high dimensional features and introduce very big optimization problems. Therefore, they are usually very slow to fit a face, which is not suitable for real-time applications.

To speed up the fitting process, Hu et al. [7] proposed a fully automatic linear algorithm to recover the shape information according to sparsely corresponded 2D facial feature points. The 3D face geometry was recovered from a frontal view input face image and then the texture was extracted from the input image directly. Being a state-of-the-art method for fitting 3DMM to sparse features, the algorithm works very efficiently and gives reasonable reconstruction results. The main restriction of this work is that it requires the input to be a frontal face image. This requirement is hard to satisfy when the face image is captured passively. In addition, the texture mapped from a single image cannot cover the whole face model and thus there will be holes after mapping the texture to the reconstructed 3D face shape.

Following the sparse feature fitting algorithm in [7], this thesis introduces a new method, which can preserve the substantial advantages and overcome the restrictions of the old method. Using Morphable Model and iterative optimization techniques, our efficient coarse-to-fine algorithm can fit a couple of arbitrary view face images to a dense 3D model of the face through an

intermediate sparse 3D point set. To obtain the better performance considering both accuracy and efficiency, the 3D Morphable Model is re-sampled with the variable sampling rate proportional to the local error made by a state-of-the-art sparse fitting algorithm.

This local quantitative error is measured by an evaluation process which compares the reconstructed shapes with the ground-truth faces. This framework is the first direct quantitative evaluation method for 3DMM based reconstruction algorithm.

After recovering the shape, the facial texture is mapped and merged from both of the images. With this method of texturing, our system does not require the input to be frontal because we can utilize the recovered camera matrix to do texture reverse mapping from images of any view. Moreover, collecting and combining texture from two sources, we can fill more parts of the face with mapped texture. The recovered 3D face model will then have the more complete and photo-realistic appearance.

The new fitting algorithm and the evaluation method are the two main contributions of the thesis. The preliminary results of this research have been reported in [8], and [9].

This thesis is organized as follows. Chapter 2 reviews the related works. The fundamental idea of 3DMM and several typical face fitting methods based on it are introduced in Chapter 3. The proposed method for quantitative evaluation for reconstruction algorithms and its results are described in Chapter 4. Chapter 5 addresses the details of the new reconstruction algorithm. In Chapter 6, the experiments and results are presented. Chapter 7 draws the conclusion and discusses future work.

CHAPTER 2

RELATED WORK

Along the line of fitting 3DMM to sparse features using multiple images, the algorithm of Zhang et al. in [10] recovered the shape from one single frontal face image among the input; after that, the images of the other views were used to refine the shape estimation and supplement the texture using minimum variance estimation. Although the other views may help improve the shape model, the quality of the estimated shape is principally determined by the first fitting with the frontal face image.

In another work [11], Park and Jain introduced an approach closely related to the framework proposed in this thesis. From two frames of a video, one of them frontal, they manually pick a set of landmark points, reconstruct the sparse 3D set of the points and then fit it to a generic model using thin plate splines [12].

This method is significantly different from ours since the method in [11] finds the dense shape model from sparse 3D points and then interpolates a generic model based on the sparse 3D point set. In our work, however, the Morphable Model is driven by priors implied in the coefficients of eigenvectors which span the PCA space learned from training faces. In this way, the new shape is constrained to be a combination of known real faces and therefore not only accurate but also more likely to be a regular human face shape. Besides, in the texture mapping step, instead of extracting texture from one image, our work estimates the texture from both images and merges them up to form the model texture. In our new approach, the restriction of using frontal image is loosened and more parts of the face textures will be reconstructed.

CHAPTER 3

3D MORPHABLE MODEL FOR FACE FITTING

3.1 3D Morphable Model

In this chapter, we will introduce the 3D Morphable Model (3DMM), the framework for representing, manipulating, and searching shapes and texture of 3D objects. This framework is the infrastructure for a series of face reconstruction algorithms including our approach for face fitting.

Based on the analysis by synthesis framework, the approach of the 3D Morphable Model presented by Blanz and Vetter in [2] proposes to model a particular human face as a linear combination of a small set of known 3D faces.

In the space spanned by 3D face examples, each face is presented by two vectors of shape and texture. The shape vector S contains the 3D location of a set of n corresponding vertices of the face:

$$S = (X_1, Y_1, Z_1, X_2, Y_2, Z_2, ..., X_n, Y_n, Z_n)^T \tag{3.1}$$

whereas the texture vector T contains the RGB color of those vertices:

$$T = (R_1, G_1, B_1, R_2, G_2, B_2, ..., R_n, G_n, B_n)^T \tag{3.2}$$

On this linear space, faces can be produced by linear combinations of shape vectors S_i and texture vectors T_i of m example faces.

$$S = \sum_{i=1}^{m} \alpha_i S_i,\ T = \sum_{i=1}^{m} \beta_i T_i \tag{3.3}$$

To reduce the dimensions of the space, principal component analysis is applied to the 3D face training database obtained from 3D laser scan. As a result, we can express the face as the combination of a small number p of

shape and texture principal components:

$$S = \bar{S} + \sum_{i=1}^{p} \alpha_i S^i,\ T = \bar{T} + \sum_{i=1}^{p} \beta_i T^i \tag{3.4}$$

where $\bar{S}$ and $\bar{T}$ are mean vectors, S^i and T^i are principal components of shape and texture respectively.

Making the assumption that human faces share similar patterns of shape and texture, and that the training data cover the range of possible human faces, we suppose that using the 3DMM, we can synthesize the face of any particular person. This supposition is the basis for 3D face reconstruction methods introduced in the next section.

3.2 3D face reconstruction based on 3DMM

Given the input as a single 2D image of a human face, we would like to find a 3D face that would produce the most similar image if it was photographed under some specific condition. Using 3DMM, we can solve that problem by finding the optimal parameters of the generalized model (principal coefficients) together with parameters of environment (illumination) and camera calibration (pose) that can synthesize the 2D image best fit with the input.

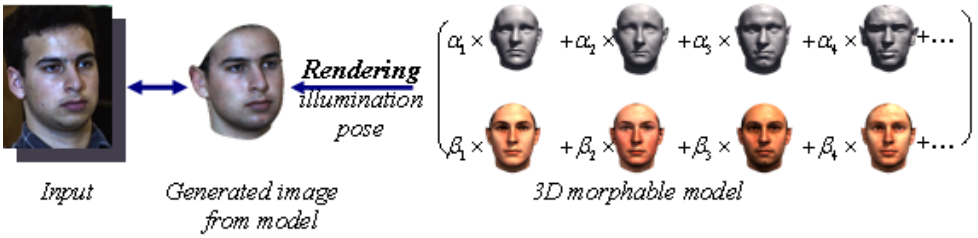

Figure 3.1: Fitting 3D Morphable Model with a single 2D image.

This scenario, illustrated in Fig. 3.1, leads us to an optimization problem where the model and environment parameters are optimized variables. The objective function of this problem would be the dissimilarity of input and synthesized image calculated on some image feature which also is minus log likelihood function. One example of the appearance features is pixel intensity

[6]. In this case, the cost function has the form of

$$C(\alpha, \beta, \rho, \iota) = -log(p(I|\theta)) \propto \sum_{x,y} (I_{model}(x, y, \alpha, \beta, \rho, \iota) - I_{input}(x, y))^2 \quad (3.5)$$

where α, β are shape and texture PCA coefficient vectors, ρ, ι are parameters for illumination and pose respectively, and θ is the long concatenation of those four vectors. This technique has proved to be effective in achieving high recognition rate and robust against different PIE (pose - illumination - expression) conditions.

In this algorithm, since the model is based on the dense correspondence set of as many as thousands of 3D vertices, the above cost function would consist of so many terms that the algorithm will be extremely computationally expensive. To obtain a more efficient fitting algorithm, in Hu et al.'s work [13], the problem is simplified by using the sparse correspondence set of feature points and estimating only the shape parameters whereas the texture will be mapped from the input. This method has proved to give a reasonable result if the input has been taken with the straight frontal pose and good illumination condition with a much alleviated computational workload. The outline of this algorthim is depicted in Fig. 3.2

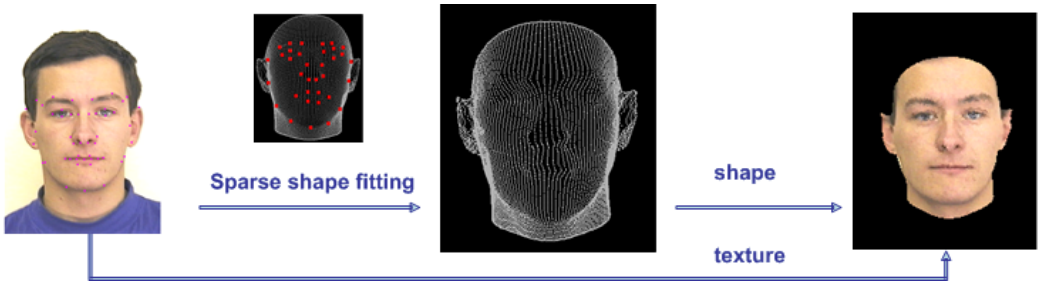

Figure 3.2: Fitting 3D Morphable Model to sparse 2D feature point of frontal face.

3.3 Improvement directions for 3DMM based algorithms

Looking at the drawbacks and limitations of the current approaches of using 3DMM for face reconstruction and finding ways to improve them, we

have come up with some ideas which would alleviate the limitations of those methods and improve the result.

One way to deal with the drawbacks of algorithms is to combine them in the hope that one method's weaknesses can be assuaged by the strength of another. For 3D object reconstruction, over the decades, epipolar geometry based methods have matured with various algorithms and techniques [14]. Using two images captured from different viewpoints, under some particular conditions, the 3D shape of an object can be reconstructed accurately and efficiently by stereopsis or other methods exploiting geometrical properties of the scene. The main drawback of these techniques is the sensitivity to noise. Besides, considering all of the 3D points in the space equally, the models reconstructed are usually lacking in completeness and correspondence. These limitations can be overcome by using Morphable Model fitting as a second phase reconstruction. In this phase, the sparse, noisy reconstructed point cloud will be corrected, normalized and aligned into correspondence frames. Our new algorithm for 3D face fitting from stereo images is based on this idea and will be described in detail in Chapter 5

Another idea to lighten the computation workload is redefining the correspondence map of the face according to the importance and complication of particular parts of the face in analysis tasks. As some parts of the face such as eyes, mouth and nose contain the most geometry and texture features of the face, we hypothesize that these parts may give the most important clues for analysis and also are the most difficult areas to analyze. This hypothesis can be proved by a local quantitative evaluation.

To assess and develop each of these two improvement ideas, it is essential to have insight into how and where the algorithms show their drawbacks. This requirement raises the need for a quantitative evaluation method for reconstruction algorithms which is proposed in the next chapter.

CHAPTER 4

A QUANTITATIVE EVALUATION FOR 3D FACE RECONSTRUCTION ALGORITHMS

4.1 Introduction

In order to demonstrate the effectiveness and evaluate the accuracy of these 3D fitting algorithms, different applications are conducted from computer graphics to face recognition. In computer graphics, the reconstructed 3D faces are rendered in different PIE and driven by MPEG Facial Animation Parameters (FAP)[15], which enriched the human computer interaction and improved the user experiences. In face-based biometrics, the reconstructed 3D faces are used to normalize or expand the probe/gallery data set under different PIE conditions, so that the test patterns are closer to the reference patterns before the matching step [7, 16]. However, all the above evaluation methods are either based on subjective experiments or indirect evaluation on face recognition, i.e., there is no information about the shape/texture error provided, which made it difficult to further analyze the reconstruction algorithm and improve the features they used. An effective quantitative evaluation is required to give us the clues for finding the strength of the algorithms, investigating their weakness and suggesting further guidance for feature selection and algorithm refinement.

In this thesis, we proposed to use signal-to-noise ratio (SNR) and error maps (EM) to quantitatively evaluate the accuracy of a 3D face reconstruction algorithm and provide its detailed performance on shape recovery. The framework of our quantitative evaluation algorithm is described in Fig. 4.1. From the ground truth 3D face database [17], we obtain the input 2D face image by projecting a 3D face onto 2D plane. This 2D face image and the extracted features will then be fed to the evaluated reconstruction system to get the reconstructed 3D face. To compare the ground truth 3D face shape and the reconstructed 3D face, they are first aligned to each other by *iterative*

closest point algorithm. The difference returned from the fitting process will then be used as the error term for calculating SNR. These measurements on all the vertices will congregate to form the error map, which provides final detail result of the evaluation process.

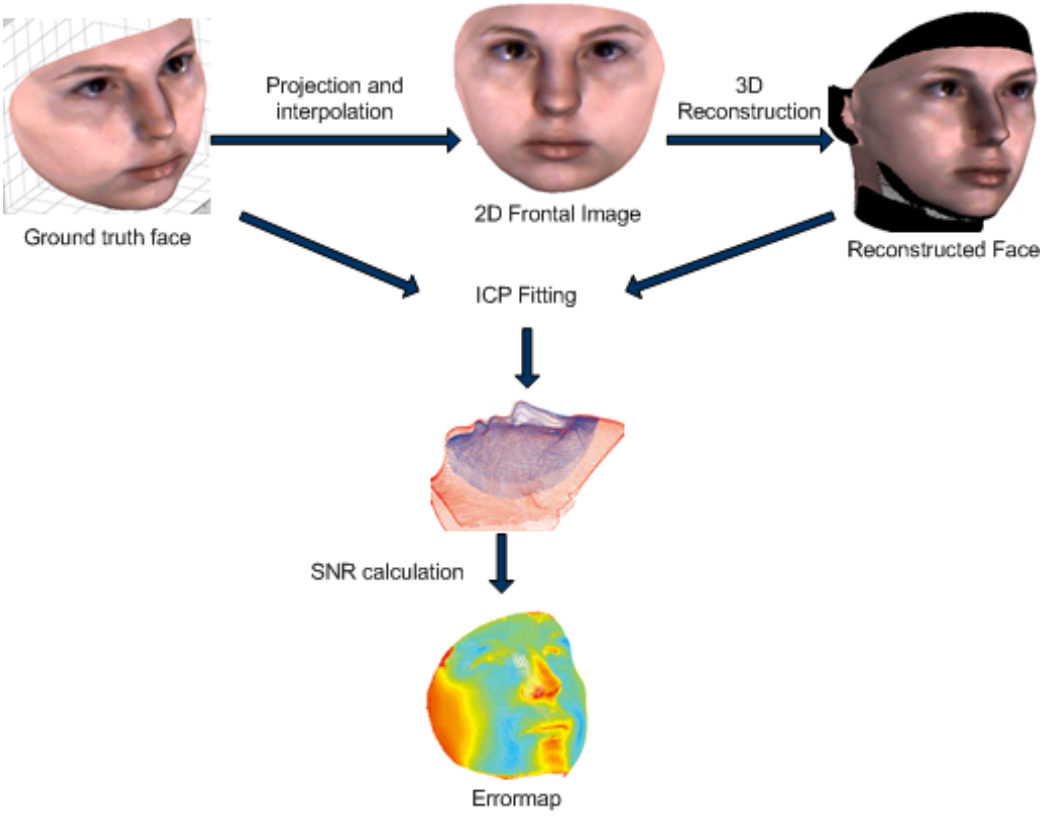

Figure 4.1: Framework of quantitative evaluation algorithm.

4.2 Direct quantitative evaluation method

The direct evaluation method is based mostly on comparing the reconstruction result with the original ground truth face. These two may be represented by different numbers of vertices and may not be in correspondence. Therefore, at the first step, we need to align the two faces by rotation and translation. *Iterative closest point algorithm* (ICP) is used for this fitting task. The ICP algorithm was introduced in [18] as an accurate and efficient method for registration of 3D shapes. The algorithm takes one shape considered as a model and another as test data, and it outputs the translation and

rotation for the data to fit with the model. With respect to the mean square error objective function, ICP was proved to always converge monotonically to the local minimum.

After aligning the faces, in order to calculate the error in terms of SNR at each vertex, we need to define the quantities for the roles of signal and noise. For noise, the obvious choice is the distance from considered point p on the original face O to the nearest point q in the reconstructed face after fitting R. This distance represents the misaligned quantity of the reconstruction at that point and is invariant to coordinate systems. The noise of each coordinate component can be expressed as the absolute differences

$$N_x^p = |p_x - q_x|, N_y^p = |p_y - q_y|, N_z^p = |p_z - q_z| \tag{4.1}$$

with $p \in O$ and q is the closest point to p in R. The combined noise at a position is defined to be the distance between two 3D points

$$N^p = \sqrt{N_x^{p2} + N_y^{p2} + N_z^{p2}} \tag{4.2}$$

Unlike the case of noise, choosing a quantity to be in the role of signal amplitude is not straightforward. Naturally, considering the mentioned definition of noise, we should use the coordinate amplitude of vertices as signals. However, this quantity depends on the frame of reference; i.e., when we change the origin or axes of coordinate system, these numbers will change. This fact would prevent the SNR from being well defined. We can address this issue by finding the coordinate system with which the amplitude of coordinate would have the smallest value. We can show that for a point $p(p_x, p_y, p_z)$ in the space, the summation of distances from it to all vertices in a set is minimized if it is the centroid $c(c_x, c_y, c_z)$ of the set.

$$\sum_{q \in O} norm(c - q) \geq \sum_{q \in O} norm(p - q) \qquad for\ all \qquad p \in \mathbf{R}^3$$

From this observation, we will choose the centroid of the face to be the origin of the coordinate system, and the maximum distance between the origin and the vertices will be considered as the amplitude of the signal.

$$S_x = \max_{p \in O}(|p_x - c_x|), S_y = \max_{p \in O}(|p_y - c_y|), S_z = \max_{p \in O}(|p_z - c_z|)$$

Signal amplitude for combined evaluation will be

$$S = \max_{p \in O} \sqrt{(c_x - p_x)^2 + (c_y - p_y)^2 (c_z - p_z)^2} \qquad (2)$$

Once the signal S and noise N are ready, SNR can be calculated at each point as

$$SNR_x^p = 20 \times log_{10}(S_x / N_x^p)$$

$$SNR_y^p = 20 \times log_{10}(S_y / N_y^p)$$

$$SNR_z^p = 20 \times log_{10}(S_z / N_z^p)$$

and combined SNR:

$$SNR^p = 20 \times log_{10}(S / N^p) \qquad (4.3)$$

4.3 Evaluation experiments and results

We applied our evaluation method on the algorithm introduced in [13]. It first conducted the face detection and 2D face alignment on the input frontal image. After that, the allocated key facial points are used to compute the 3D shape coefficients of the shape Morphable Model. The result of this step is the 3D face shape represented by a combination of principal components. Hence, it has the same format of those components which includes 8955 3D points, covering the whole face region as the green shapes in Fig. 4.2.

The ground truth used to evaluate the above algorithm is the IFP 3D face database [17], which consists of 500 3D faces obtained by laser scanner. More details about this database are given in Appendix A. Each face in ground truth is presented by 33,420 vertices with dense correspondences, covering the frontal part of the face as depicted as red shapes in Fig. 4.2.

In the first step of our process, for each face in the database, the frontal image is created by projecting the 3D face to the frontal plane. The resulting 2D face image is fed into the reconstruction system as an input. The output of this process is a 3D shape in the same scale but with different number of vertices. The vertices in the output 3D face are not in correspondence with the ones in the original ground truth data. So we need to find the best fit for them and then calculate the difference between corresponding vertices, where ICP will be applied for the solution.

In our case, the reconstructed face covers a larger area than the original face. Thus, we will use it as model and the original face will act as fitted data. Since ICP cannot guarantee global minimum, we first manually find the transformation parameters which give every pair of faces a relatively near-to-optimal position. The ICP iteration is then employed to get the fitting result (rotation and translation parameters) together with the square error on each vertex of the original face. A sample of starting position and fitting result is depicted in Fig. 4.2(a) and 4.2(b), respectively.

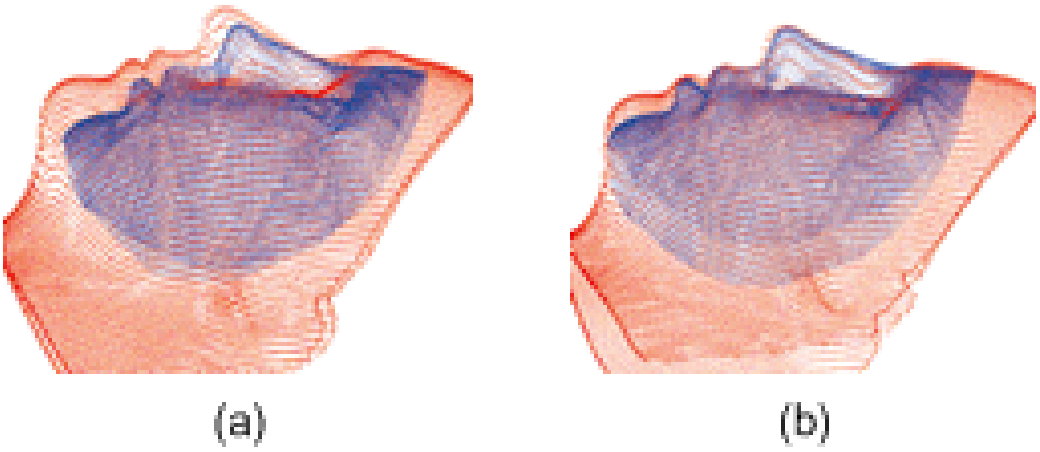

Figure 4.2: Fitting result. Original face is in green, reconstructed shape is in red. (a) Initial position, (b) fitted shapes.

The experiment procedure is applied on 50 faces in IFP database. After fitting, the error maps are calculated as in Equation 4.3. The mean of combined SNR is 28.96; for the x component it is 31.62, for the y component 38.64 and for the z component 30.00. The error maps are shown in Fig. 4.3.

We can observe on the map that the most significant inaccuracy was made on the nose tip, chin and cheek areas of the face. Moreover, the error of the z component is the most considerable. This can be explained by the nature of the reconstruction algorithm, which uses 2D facial points on the frontal face so that the z component is determined purely by the statistical information embedded the 3D Morphable Model.

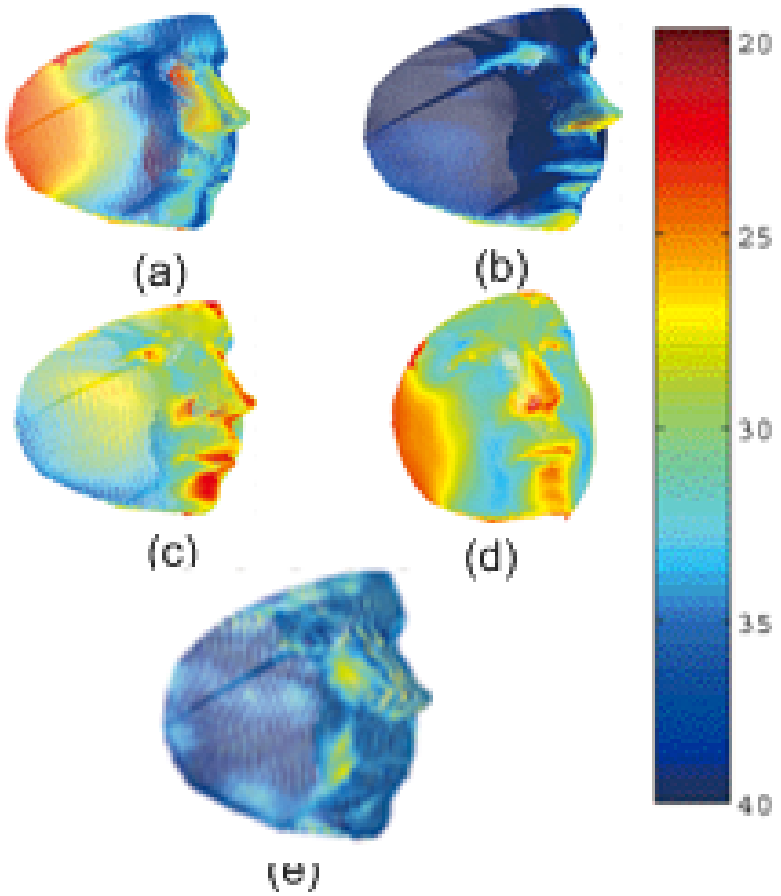

Figure 4.3: Error map of (a) x components, (b) y components, (c) z components, (d) combined signal, (e) ideal reconstruction in SNRDB.

4.4 Upper bound of reconstruction

The error we got from fitting the reconstructed and ground truth 3D faces comes from both the reconstruction inaccuracy and the noise of signal sampling and shape fitting. To assess the significance of those two components, we directly converted the ground truth faces to the format of the output to simulate an ideal reconstruction. Comparing this ideal reconstruction result with the ground truth gives us the error caused only by sampling and fitting but not by reconstruction. The error map from this comparison is shown in Fig. 4.3e and the combined SNR is 34.21.

The error map shows that the noise made by other causes is relatively uniform over the face. This fact proves that the variety of error at different parts of the face as in Fig. 4.3 comes from the reconstruction. This variety will provide us precious information about the strong and weak points of the

considered algorithm.

The ideal reconstruction error rate not only shows the underlying noise, but also provides a very important landmark, which can be used as the upper bound of the reconstruction performance that any reconstruction algorithms can reach.

4.5 Discussion

The evaluation result on the selected reconstruction algorithm [7] shows that the error concentrates at some particular parts of faces such as nose tip, chin and cheek.

The reason for the inaccuracy of height at nose and chin areas is that, in the reconstruction algorithm we evaluated, the depth information is deduced from the width/height information in the 2D face according to the statistical model, whose performance is limited by the size of the training set. The reason for the inaccurate cheek is that the 2D facial contour actually is formed by projection, which is not well defined in the 3D face, so their corresponding points cannot be precisely located.

In summary, the inaccuracy comes from the limited training data and the lack of control points at salient and complicated areas. These observations suggest new strategies of 2D control points configuration and building larger database or specific training set for different ethnicity/gender groups. These suggestions will be applied in the design of a novel algorithm for fitting 3D faces to couples of stereo images, which is introduced in the next chapter.

CHAPTER 5

3D FACE RECONSTRUCTION FROM STEREO IMAGES

Our reconstruction algorithm works in a coarse-to-fine fashion. First, from a collection of salient points located on two input face images, a sparse set of corresponding 3D points is reconstructed. After that, these reconstructed 3D points will be used to build a dense 3DMM model of the facial shape. Finally, the texture from both input images will be mapped and merged altogether on the reference frame of the Morphable Model. The entire process will generate a detailed, textured 3D face model. The framework for our method is depicted in Fig. 5.1.

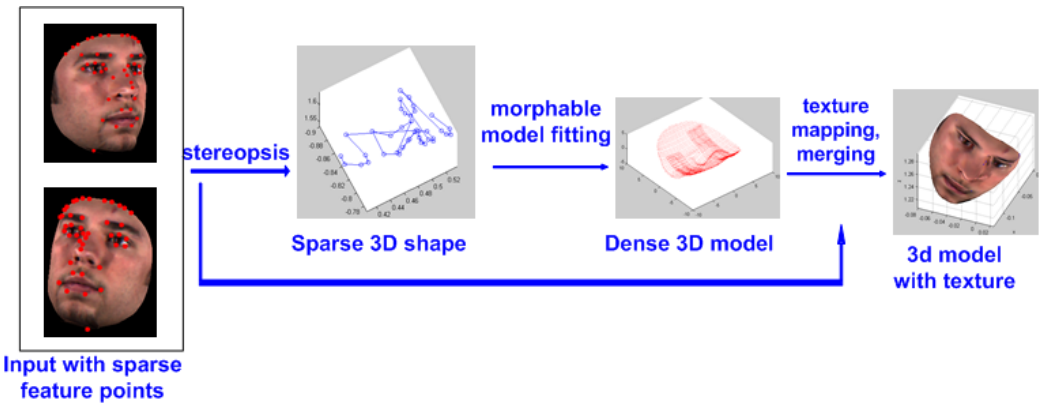

Figure 5.1: Coarse-to-fine 3D face reconstruction framework.

5.1 3D sparse shape reconstruction

In order to have the stereopsis shape reconstruction work, we need to make an assumption that the cameras are calibrated. For the studio data, this assumption is normally true; for other cases, self calibration techniques can be applied to find the intrinsic camera parameters.

To reconstruct a sparse set of 3D feature points, we first need to locate enough facial feature points and find their correspondences on the two input images. This can be done automatically using a face alignment algorithm such as the one in [19]. With the corresponding feature points on the two calibrated images, the 3D feature points are reconstructed by triangulation. A sample input couple and its reconstructed sparse shape are depicted in Fig. 5.2.

Figure 5.2: Sparse shape reconstruction from two different view images.

5.2 3D dense shape model fitting

In this step, we fit the sparse set of 3D points to a dense face model by searching in the linear space spanned by a set of training 3D faces. The training data is projected into the PCA subspace, where the 3DMM is described by a vector of principal component coefficients.

Specifically, in our framework, a new 3D dense face shape is represented in the linear space as

$$S' = \bar{S} + P\overrightarrow{\alpha} \tag{5.1}$$

where $\bar{S}$ is the average shape, P is the matrix of shape principal components and $\overrightarrow{\alpha}$ is the vector of the corresponding principal component coefficients.

Now, let s', $\bar{s}$ and p be the sub-vectors of S', $\bar{S}$ and P, respectively, corresponding to the obtained sparse 3D point set. In the subspace of sparse shapes, a new sub-face is constructed by

$$s' = \bar{s} + p\overrightarrow{\alpha} \tag{5.2}$$

Considering the sparse shape obtained from the previous step, s'', its registration with the new shape is represented by

$$s'' = cRs' + t \tag{5.3}$$

where R is the rotation matrix, c is the scale factor and t is the translation vector.

Given s'', our objective is to recover S' by finding $\overrightarrow{\alpha}$ from Equations 5.2 and 5.3. To solve this problem, we introduce an iterative procedure inspired by the algorithm for fitting 2D feature points [7] with the following steps.

Initialization: Assign $\bar{s}$ to s'.

Step 1: Apply the procrustes analysis to solve Equation 5.3. In this step, the R, c, and t that would fit s' and s'' the best are found. In our algorithm, we use the method of finding the eigenvector of the correlation matrix corresponding to the least eigenvalue for the procrustes analysis.

Step 2: Find α by solving Equation 5.2 as a linear least-squares problem with the constraints $\alpha \leq r$ where r is some constant threshold vector. These inequality constraints emerge as the restrictions to keep the new face having regular human face shape.

In order to find a suitable threshold r, we employ a commonly used assumption that the facial shape has normal priors with a probability density function of

$$p(\alpha) \sim e^{\frac{1}{2}\sum_i \frac{\alpha_i^2}{\sigma_i^2}} \tag{5.4}$$

This leads us to the choice for constraint thresholds proportional to the standard deviation of the representation of the training data in the principal component space: $r_i = \lambda\sigma_i$ (λ is a constant). Our experiments show that this choice for threshold gives better results than using the eigenvalues as in [7]. Solving 5.1 with these constraints, the shape principal coefficients will be found as

$$\alpha = (p^T p + \lambda \mathrm{diag}(r^2)^{-1})^{-1} p^T (s' - \bar{s}) \tag{5.5}$$

After initialization, Steps 1 and 2 are alternatively repeated in each iteration. In our experiments, the process converges in at most 10 rounds. The fitting result of a sample face is depicted in Fig. 5.3

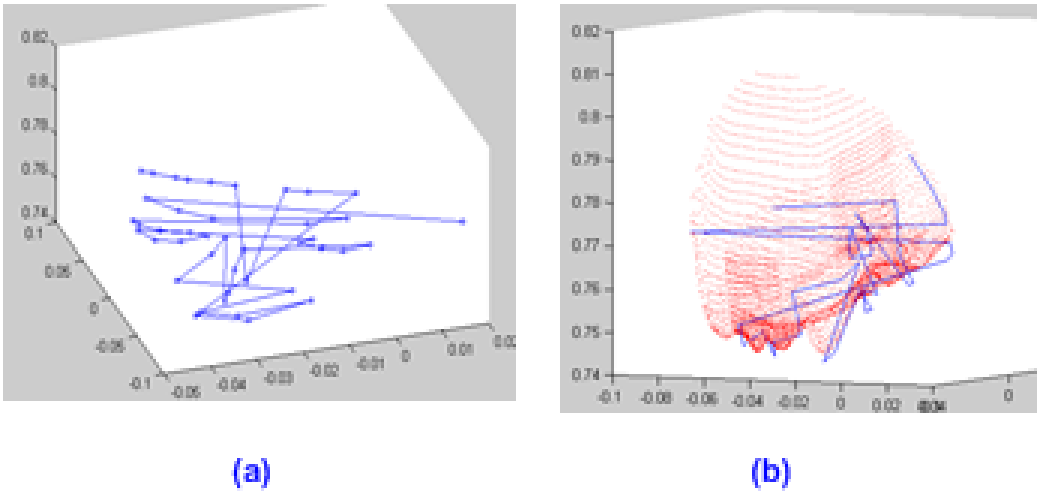

Figure 5.3: Fitting sparse set of 3D points to 3D Morphable Model.

5.3 Texture mapping

After reconstructing the shape, the next step is to recover the texture of the face model. In this step, the recovered shape is projected onto both input images to find texture coordinates of model's vertices. At each viewpoint, a depth buffer is used to check the visibility of the vertices of the face; only the visible ones are textured by inverse mapping. In the areas of the face where two extracted textures overlap, the color of a vertex will be the average of two texture values. Figure 5.4 illustrates the texture obtained from two images and the merged texture.

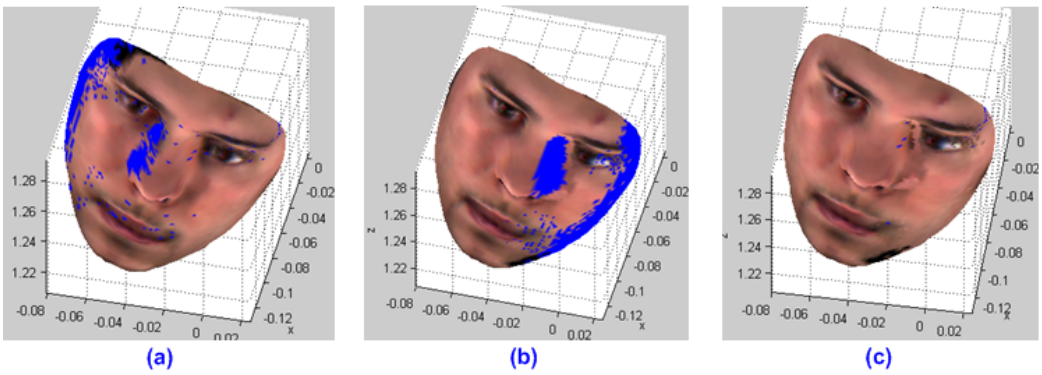

Figure 5.4: Texture mapping and merging. (a), (b) Texture mapped from two images. (c) Combined texture. The blue areas indicate the locations where texture is missing.

CHAPTER 6

EXPERIMENTAL RESULT

6.1 Experiment configuration

To test our algorithm, we again conduct experiments on the IFP 3D Face Database [17]. The evaluation follows the procedure introduced in Chapter 4 with 5-fold cross validation.

During the training process, we project 400 faces in 4 training folds into a linear space with PCA. The correspondence of the vertices is originally defined on a uniformly sampled cylindrical mask, as depicted in Fig. 6.1(a). As discussed in Chapter 4, the reconstruction inaccuracy is the most substantial in the areas near eyes, noses, and lips where there are lot of details. Moreover, these parts are significant areas for face recognition and face animation. To reduce errors on these crucial areas, before performing PCA, we propose to resample the mask model, assigning higher weights to those parts of the face mask. The resampling process also helps reduce the number of vertices of the model to a reasonable number (about 6000 in our experiments), which provides a good compromise between speed and accuracy. The new weighted reference frame is plotted in Fig. 6.1(b).

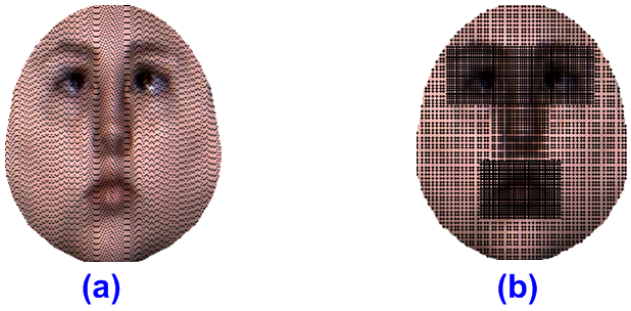

Figure 6.1: Original (a) and resampled (b) reference masks: More samples are given at important and complicated areas.

For each of the 100 3D face models in the testing fold, four pairs of syn-

thesized images are generated by simulating pinhole cameras. Each image pair enjoys a particular setup of the direction and relative position of the camera and the face. These various setups are summarized in Table 6.1. The direction is represented by trinities of yaw-pitch-roll angles. The position of each camera is arranged so that the principal axis points to the center of the face from a given distance. The feature points for sparse reconstruction are located by projecting the corresponding 3D points to the image planes.

By collecting the image pairs having the same setups, we form four test sets, each with 100 pairs of features and images. The procedure for fitting faces as addressed in Chapter 5 is applied on the test sets. The results of the experiments will be shown in the next section.

Table 6.1: Setups for image pairs.

Setups	Image1			Image2		
	Angles (degrees)	Distance (cm)	Sample image	Angles (degrees)	Distance (cm)	Sample image
1: one frontal, one other view	0, 0, 0	150		30, 0, 0	100	
2: two symmetrically non-frontal views	-30, 0, 0	150		30, 0, 0	100	
3: various yaw-pich-roll angles	-20, -10, 10	150		30, 10, 0	120	
4: with an extreme yaw angle	-20, -10, 10	120		60, 0, 0	120	

6.2 Results and evaluation

Figure 6.2 illustrates the reconstruction results of our algorithm on the corresponding samples on Table 6.1. Subjectively assessing, the face shapes and textures are accurately reconstructed. The texture is reconstructed almost completely, which covers both the frontal and side areas of the face without holes and gives a realistic look. Compared with related texture mapping based methods such as [7], our algorithm generates more complete face texture due to the use of the texture from all input images, as shown in Fig. 6.3.

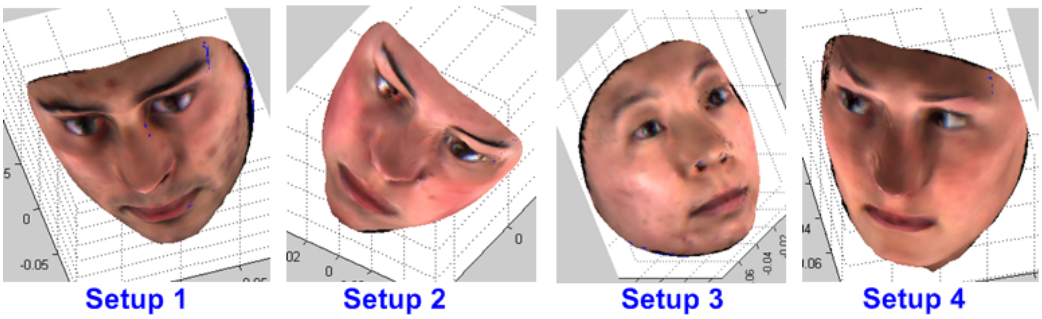

Figure 6.2: Rendered reconstruction results.

Figure 6.3: Comparison of fitting results of Hu's and our algorithm with ground truth.

The accuracy of the reconstruction is also quantitatively evaluated by measuring the disparity between the synthetic shape and the ground truth shape used to generate input images. We align those couples by Procrustes transformation and measure the distance between the corresponding vertices. The mean RMS error of all the setups is 2.3 mm. Since most previous 3DMM based algorithms are only evaluated indirectly by recognition accuracy [20], we are not able to compare our algorithm with them. However, we have done evaluation on the work in [7]. We will compare our method with this state-of-the-art work.

The graph of mean SNR measured in dB of our algorithm on four setups compared with Hu's algorithm is shown in Fig. 6.4. The error map for the local SNRDB is depicted in Fig. 6.5.

From the SNR graph and error maps, we can see that our method consistently performs better in all four test sets. More importantly, the error also decreases significantly in the important areas of the faces. With the proposed resampled correspondence mask, those parts are the most accurately reconstructed.

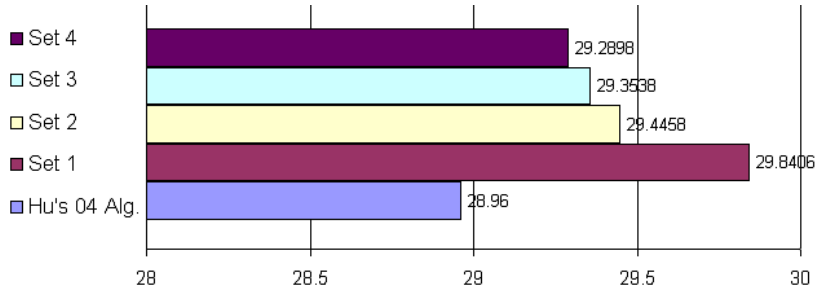

Figure 6.4: SNRDB of our algorithm on four setups and Hu's algorithm.

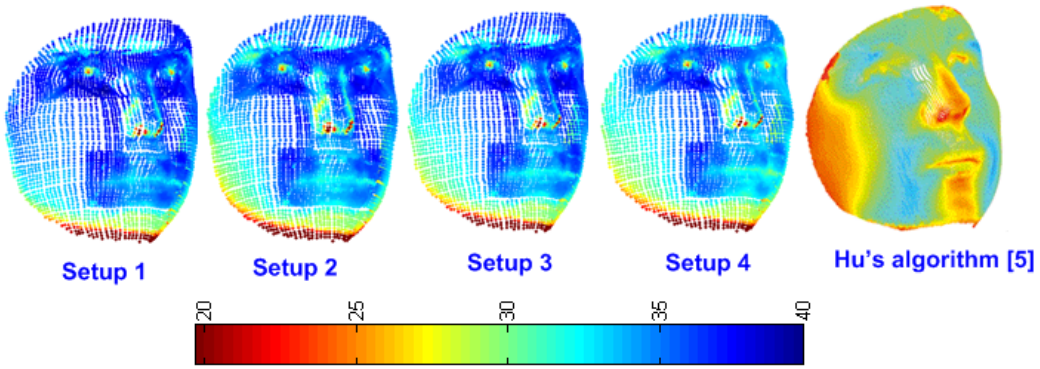

Figure 6.5: Error map in SNRDB of our algorithm on four setups and Hu's algorithm drawn with colormap.

We implement our algorithm in Matlab, and the time used to recover model shape and texture is 5.2 seconds on a 2 GHz Core Duo CPU laptop computer. Our speed is comparable to that of Hu's algorithm [7] and is much higher than those of dense feature based algorithms [1, 5]. The speeds of the various algorithms are compared in Table 6.2.

Table 6.2: Speed comparison of our algorithm with related works.

Algorithm	Running time	Processor
Vetter's 99 (SNO)	4.5 mins	2.0GHz
Vetter's 2003 (ICIA)	30 secs	2.8GHz
Vetter's 2005 (MFF)	70 secs	3.0GHz
Hu's 04	~5 secs	1.3 GHz
Ponce's 2007	>20 mins	N/A
Our method	5.2 secs	2.0GHz

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this thesis, we reviewed the family of 3D face reconstruction algorithms based on 3DMM. We proposed and experimented upon a new framework for quantitatively assess the performance of these algorithms. The evaluation not only showed the accuracy level of current algorithms but also provided information about local error they made, which is an important clue for feature selection and algorithm design.

The thesis also proposed a novel algorithm for reconstructing the 3D shape and texture of human faces from two stereo images captured from calibrated cameras. The features were selected based on the evaluation of a state-of-the-art algorithm. The algorithm accepts the input of virtually any pose of the face. The restriction of using frontal view of the compared algorithm is exempted and the result is more complete and photo realistic.

Subjective and quantitative evaluations have shown that the proposed reconstruction algorithm is efficient and accurate with the input of stereo images. With these promising results, our future work will generalize the current algorithm to the scenarios with more input images, and integrate automatic feature point localization and camera self-calibration. We believe our work will be useful for building a fully automatic system of 3D face reconstruction, recognition and animation in practical applications.

APPENDIX A

IFP 3D FACES DATABASE

The systems introduced in this thesis used IFP 3D Faces database as the main source of training and evaluation data. This database is created and post-processed in our laboratory at the University of Illinois. It is a large scale 3D face database with dense correspondence. *Large scale* means that the number of subjects in the database is more than 400. *3D face* means that we provide both the texture and shape of human faces, which are also balanced in gender and race. *Dense correspondence* means that the key facial points with semantic meanings are carefully labeled and aligned among different faces, which can be used for a broad range of face analysis tasks [17].

A.1 Production procedure

The database has been created by 3D laser scan and a multiple-step post-processing procedure. For the acquisition of the raw 3D face data, 500 subjects were invited for the 3D face scanning. For each subject, a 3D human face was first scanned by a Cyberware scanner, which output the raw data including the head shape and texture map. Then 65 pre-defined key facial points, such as eye/mouth corners and nose tip, were labeled on the texture map manually. According to these key points and the reference mean face model, the face region was cropped out and the dense triangular mesh shape model was created by interpolation so that the vertices in different face models corresponded with semantic meaning. Finally, the poses of the 3D faces were normalized so that their 3D coordinates were in the same reference frame. Some samples of the database are depicted in Fig. A.1.

Figure A.1: Samples of the 3D face after post-processing.

A.2 Data format

For each face, together with the shape and texture of the normalized 3D face model, the information of the subject, including age, gender and ethic group is available. The shape and texture map of the raw data are stored separately, occupying about 3 GB of disk space. After post-processing, there are 33,420 vertices/points for each face, which is about 400 MB in file size.

REFERENCES

[1] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, 2007, pp. 1–8.

[2] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Siggraph 1999, Computer Graphics Proceedings*, A. Rockwood, Ed. Los Angeles: Addison Wesley Longman, 1999, pp. 187–194.

[3] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Intl. Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, November 1992.

[4] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 690–706, 1999.

[5] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3d morphable model," in *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV '03)*. Washington, DC, USA: IEEE Computer Society, pp. 59–66.

[6] S. Romdhani and T. Vetter, "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005)*, vol. 2, 2005, pp. 986–993.

[7] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang, "Automatic 3d reconstruction for face recognition," in *International Conference on Automatic Face and Gesture Recognition (FGR2004)*, Nagoya, Japan, 2004, pp. 843–848.

[8] V. Le, Y. Hu, and T. Huang, "A quantitative evaluation for 3d face reconstruction algorithms," in *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1269–1272.

[9] V. Le, H. Tang, L. Cao, and T. Huang, "Accurate and efficient reconstruction of 3d faces from stereo images," submitted to *IEEE International Conference on Image Processing 2010 (ICIP 2010)*, 2010.

[10] Z. Zhang, Y. Hu, T. Yu, and T. Huang, "Minimum variance estimation of 3d face shape from multi-view," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR '06)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 547–552.

[11] U. Park and A. K. Jain, "3d face reconstruction from stereo video," in *Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision (CRV '06)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 41–48.

[12] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 6, pp. 567–585, 1989.

[13] D. Jiang, Y. Hu, and S. Yan, "Efficient 3d reconstruction for face recognition," *Journal of Patern Recogognition, Special Issue on Image Understanding for Digital Photographs (PR2005)*, vol. 38, pp. 787–798, 2005.

[14] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 519–528.

[15] H. Tang, Y. Hu, Y. Fu, M. Hasegawa-Johnson, and T. S. Huang, "Real-time conversion from a single 2d face image to a 3d text-driven emotive audio-visual avatar," in *IEEE International Conference on Multimedia & Expo (ICME 2008)*, 2008, pp. 1205–1208.

[16] V. Blanz, S. Romdhani, and T. Vetter, "Face identification across different poses and illuminations with a 3d morphable model," in *International Conference on Automatic Face and Gesture 2002 (AFGR'02)*, 2002, pp. 192–197.

[17] Y. Hu, Z. Zhang, X. Xu, Y. Fu, and T. Huang, "Building large scale 3d face database for face analysis," in *Multimedia Content Analysis and Mining, International Workshop (MCAM07)*, 2007, pp. 343–350.

[18] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Transaction on Pattern Analysis Machine Intelligent*, vol. 14, pp. 239–256, 1992.

[19] D. Cristinacce, T. Cootes, and I. Scott, "A multi-stage approach to facial feature detection," in *British Machine Vision Conference (BMVC'04)*, 2004, pp. 277–286.

[20] P. Phillips et al., "Face recognition vendor test 2002," Technical Report NISTIR 6965, National Institute of Standards and Technology, 2003.